Continuous, Discrete and Limited Dependent Variable Regressions with Application to Genetic Association of Body Mass Index and Obesity

J.H. Zhao[1], S. Li[1], J.A. Luan[1], Q. Tan[2], E. Wheeler[3], S. Debenham[4], M. Inouye[2], P. Deloukas[3], M. Sandhu[4], I. Barroso[3], R. McGinnis[3], R. Loos[1], N.J. Wareham[1]

[1]MRC Epidemiology Unit, Cambridge, UK, [2]Odense University Hospital, Denmark, [3]The Wellcome Trust Sanger Institute, Hinxton, UK; [4]Department of Public Health & Primary Care, University of Cambridge, UK

1. Abstract

Background Although there have been a lot of discussions over the advantages and disadvantages of regression analysis using discrete versus continuous dependent outcome, both are widely used in biomedical research. Discussion over synthesis of information from both analyses is rare.

Motivation Our revisit to the problem is motivated by our genome-wide association study (GWAS) of body mass index (BMI, weight (kg)/height (m)2) and obesity (BMI ≥ 30) based on Affymetrix 500k GeneChips and ~3,500 individuals in a case-cohort design from the European Prospective Investigation in Cancer (EPIC) Norfolk cohort.

Methods We attempt to infer genotype-phenotype relationship from sub-cohort, case-only, and case-control data by regression analyses. The impact of the dichotomisation relative to BMI as a continuous trait in the study of association with genetic polymorphisms is therefore characterised.

Findings When applied to rs9939609 in the FTO gene, logistic regression gives the most significant evidence of association, followed by linear regression in the sub-cohort. However, BMI in obese individuals provides no such evidence. The utility of continuous measurement in obese individuals and therefore the synthesis of results may be hampered by power and other complications.

Conclusion Our work serves as a good reminder for use of continuous and discrete measurement in GWAS. We discuss related issues and give some recommendations.

4 Method

- Four types of regression models have been considered
  - Linear regression of the subcohort data
  - Logistic regression of subcohort and obesity-subcohort data
  - Truncated regression of BMI in obesity cases
  - Gamma regression of BMI in obesity cases
- Meta-analysis of BMI in cohort sample and obesity cases versus logistic regression of cases and controls
- All four regression analyses are carried out with SAS and applied to rs9939609 in the FTO gene. Samples from both stages are used.

5 Results

Table 1 Parameter estimates from linear, logistic, truncated and gamma regressions at stage one

Model*	Sample	*N*	β	$SE(\beta)$	z/t	*P*
Logit	Case-cohort	3464	0.205	0.051	4.03	< 0.0001
	Sub-cohort	2364	0.245	0.082	2.98	0.003
LR	Sub-cohort	2363	0.382	0.117	3.26	0.001
	Non-sub-cohort cases	1100	-0.191	0.128	-1.49	0.14
TR	All cases	1470	-8.01	15.88	-0.5	0.61
	Non-sub-cohort cases	1097	-7.16	9.3	-0.77	0.44
GR	All cases	1474	0.0001	0.0001	1.04	0.3
	Non-sub-cohort cases	1100	0.0002	1	1.49	0.14

*Logit=logistic regression, LR=linear regression, TR=truncated regression, R=gamma regression

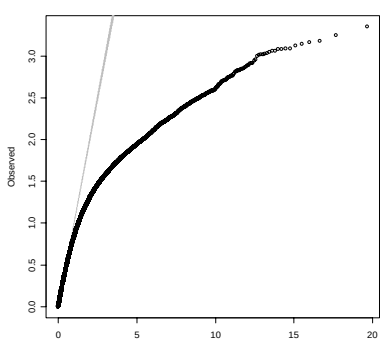
mean/standard deviation (26.4+/- 3.9) and range 16.1~47.6.

Table 2 Parameter estimates from linear, logistic, truncated & gamma regressions at stage two

Model*	Sample	*N*	β	$SE(\beta)$	z/t	*P*
Logit	Case-cohort	3760	0.201	0.05	4.05	<0.001
	Sub-cohort	2504	0.32	0.183	1.75	0.08
LR	Sub-cohort	2504	0.213	0.115	1.85	0.06
	Non-sub-cohort cases	1256	0.241	0.126	1.9	0.06
TR	All cases	1290	5475	0.02	277783	<.0001
	Non-sub-cohort cases	1256	705	332	2.12	0.03
GR	All cases	1290	-0.0002	0.0001	-2.06	0.04
	Non-sub-cohort cases	1256	-0.0002	0.0001	-1.9	0.06

The means (standard deviations) of BMI at stage 2 for the three genotypes are 28.2 (4.7), 28.6 (4.9), 29.3 (5.1) for all samples, 26.2 (3.8), 26.4 (4.0), 26.7 (4.1) for the subcohort sample and 32.9 (2.9), 33.0 (3.2), 33.4 (3.5) for the non-subcohort sample.

Fig. 2 Figure 2. Q-Q plot of p values (Chromosome 16) from PROC QLIM



2 Motivation

Our motivation to revisit the problem of discrete versus continuous outcome comes from our genome-wide association study (GWAS) of obesity, based on the European Prospective Investigation in Cancer (EPIC) Norfolk cohort (http://www.srl.cam.ac.uk/epic). Our primary outcome is obesity, defined to be body mass index (BMI, weight (kg)/height (m)2) greater than or equal to 30. The analyses of interest are:

1. Obesity, which is binary (non-obese and obese) and can be analysed through logistic regression, or equivalently, the Cochran-Armitage trend test.
2. BMI as a continuous trait, for which it is most appropriate to conduct linear regression analysis in the sub-cohort sample.
3. BMI in obese individuals, which should contain useful information. However, it is undesirable to use linear regression to examine the phenotype-genotype association between BMI in the obese individuals and SNPs as they are selected.
4. The synthesis of evidence from above models and meta-analysis with other cohorts. It is somewhat uncertain this could be done properly.

3 EPIC-Norfolk GWAS

We used a case-cohort design (Prentice 1986) such that the control sample is a proportion (about 10%) of individuals randomly sampled from the whole EPIC-Norfolk cohort, whereas the cases are from the whole EPIC-Norfolk cohort. We use Affymetrix 500k and Illumina 317k GeneChips to measure single nucleotide polymorphisms (SNPs).

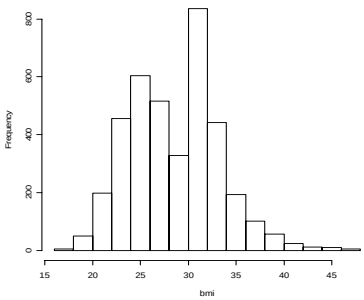


Figure 1. The BMI distribution in the EPIC-Norfolk study of obesity. The case-cohort sample is a combination of the sub-cohort sample and case sample which is truncated from the whole EPIC cohort at BMI=30

6 Simulation

The behaviour of analyses are examined through computer simulation. The model is such that the observed quantitative trait (y) is a function of the quantitative trait locus (QTL): $y = \mu + g + \varepsilon$, where μ is the overall mean, g is the putative QTL and ε is the residual. The QTL is analogous to a SNP, which is biallelic with minor allele frequency (MAF) p_1 . Depending on the number of minor allele(s), there are three possible values 0, 1, 2 for g under additive model, but there are only two possible values 0, 1 under dominant model showing absence and presence of the minor allele, or showing not to be or to be homozygous with minor alleles under recessive models.

Linear regression using only the case sample will generally give biased results, with regression coefficient and the standard error tending to be underestimated. The truncated regression will give more appropriate estimate of the regression coefficient but have larger standard error compared to the complete data due to reduction in precision of the parameter estimates from smaller subset of individuals with selected values. The regression coefficient may not be comparable to that in linear regression, but the precision of the regression parameter from logistic regression is comparable to that from the ordinary least squares.

7 Discussion

- Our analysis of rs9939609 in the FTO gene suggests that synthesis of information from cohort and case samples may not necessarily be very useful. Most likely the latter requires considerably larger sample size for a significant variation between genotypes to be seen, and the direction of association may be in the opposite direction. It would be helpful to perform descriptive analysis first for such data.
- This preliminary investigation serves as a good reminder for practitioners actively engaging in genetic association studies on issues which may arise. In contrast to the large body of literature which separates the two schools of analyses with discrete and continuous outcomes, together with discarding data in the discrete variable analysis we see they should be treated and represented in a coherent manner.

We can use the idea to other measurement. The log-likelihood function for blood pressures of the hypertensive individuals is then

$$l = \ln \int_{(L_1 - x_1\beta_1)/\sigma_1}^{\infty} \int_{(L_2 - x_2\beta_2)/\sigma_2}^{\infty} e^{-(u^2+v^2-2\rho\sigma_1\sigma_2)/2} // [2\pi(1-\rho^2)^{1/2}] du dv,$$

where x_1, x_2 represent blood pressures and L_1, L_2 are the thresholds for being hypertensive (SAS Institute Inc 2004). When applied to five SNPs reported by WTCCC, generalised estimating equations of blood pressures and bivariate truncated regression reveal no statistical significance at 0.05 level, but two with $p<0.05$ under binomial model, suggesting more complication.